Student Sliced Inverse Regression

Alessandro Chiancone, Florence Forbes, Stéphane Girard

Student Sliced Inverse Regression

Alessandro Chiancone[a,b,c,*], Florence Forbes[a], Stéphane Girard[a]

[a] *Inria Grenoble Rhône-Alpes & LJK, team Mistis, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.*
[b] *GIPSA-Lab, Grenoble INP, Saint Martin d'Hères, France.*
[c] *Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, A-8010 Graz, Austria.*

Abstract

Sliced Inverse Regression (SIR) has been extensively used to reduce the dimension of the predictor space before performing regression. SIR is originally a model free method but it has been shown to actually correspond to the maximum likelihood of an inverse regression model with Gaussian errors. This intrinsic Gaussianity of standard SIR may explain its high sensitivity to outliers as observed in a number of studies. To improve robustness, the inverse regression formulation of SIR is therefore extended to non-Gaussian errors with heavy-tailed distributions. Considering Student distributed errors it is shown that the inverse regression remains tractable via an Expectation-Maximization (EM) algorithm. The algorithm is outlined and tested in the presence of outliers, both in simulated and real data, showing improved results in comparison to a number of other existing approaches.

Keywords: Dimension reduction, Inverse regression, Outliers, Robust estimation, Generalized Student distribution.

*Corresponding Author
Email address: al.chiancone@gmail.com (Alessandro Chiancone)

1. Introduction

Let us consider a regression setting where the goal is to estimate the relationship between a univariate response variable Y and a predictor $\mathbf{X}$. When the dimension p of the predictor space is 1 or 2, a simple 2D or 3D plot can visually reveal the relationship and can be useful to determine the regression strategy to be used. If p becomes large such an approach is not feasible. A possibility to overcome problems arising in the context of regression is to make the assumption that the response variable does not depend on the whole predictor space but just on a projection of $\mathbf{X}$ onto a subspace of smaller dimension. Such a dimensionality reduction leads to the concept of sufficient dimension reduction and to that of central subspace [1]. The central subspace is the intersection of all dimension-reduction subspaces (d.r.s.). A subspace S is a d.r.s. if Y is independent of $\mathbf{X}$ given $\mathbf{P}_S\mathbf{X}$, where $\mathbf{P}_S$ is the orthogonal projection onto S. In other words, all the information carried by the predictors $\mathbf{X}$ on Y can be compressed in $\mathbf{P}_S\mathbf{X}$. It has been shown under weak assumptions that the intersection of all d.r.s., and therefore the central subspace, is itself a d.r.s. [2]. It is of particular interest to develop methods to estimate the central subspace as once it is identified, the regression problem can be solved equivalently using the lower-dimensional representation $\mathbf{P}_S\mathbf{X}$ of $\mathbf{X}$ in the subspace.

Among methods that lead to an estimation of the central subspace, Sliced Inverse Regression (SIR) [3] is one of the most popular. SIR is a semiparametric method assuming that the link function depends on d linear combinations of the predictors and a random error independent of $\mathbf{X}$: $Y = f(\beta_1^T\mathbf{X}, \ldots, \beta_d^T\mathbf{X}, \epsilon)$. When this model holds, the projection of $\mathbf{X}$ onto the space spanned by the vectors $\{\beta_i, i = 1, \ldots, d\}$ captures all the information about Y. In addition, [3] shows that a basis of this space can be recovered using an inverse regression strategy provided that the so called *linearity condition* holds. It has been shown that the *linearity condition* is satisfied as soon as $\mathbf{X}$ is elliptically distributed. Moreover, this condition approximately holds in high-dimensional datasets, see [4]. However, solutions have been

proposed to deal with non elliptical distributed predictors and to overcome the *linearity condition* limitation [5, 6, 7].

The inverse regression approach to dimensionality reduction gained then rapid attention [8] and was generalized in [9] which shows the link between the axes spanning the central subspace and an inverse regression problem with Gaussian distributed errors. More specifically, in [10, 9], it appears that, for a Gaussian error term and under appropriate conditions, the SIR estimator can be recovered as the maximum likelihood estimator of the parameters of an inverse regression model. In other words, although SIR is originally a model free method, the standard SIR estimates are shown to correspond to maximum likelihood estimators for a Gaussian inverse regression model. It is then not surprising that SIR has been observed, *e.g.* in [11], to be at best under normality and that its performance may degrade otherwise. Indeed, the Gaussian distribution is known to have tails too light to properly accommodate extreme values. In particular, [12] observes that SIR was highly sensitive to outliers, with additional studies, evidence and analysis given in [13]. To downweight this sensitivity, robust versions of SIR have been proposed, mainly starting from the standard *model free* estimators and trying to make them more resistant to outliers. Typically, in [14] classical estimators are replaced by high breakdown robust estimators and, recently in [15] two approaches are built: a weighted version of SIR and a solution based on the intra slice multivariate median estimator.

As an alternative, we propose to rather exploit the inverse regression formulation of SIR [10, 9]. A new error term modeled by a multivariate Student distribution [16] is introduced. Among the elliptically contoured distributions, the multivariate Student is a natural generalization of the multivariate Gaussian but its heavy tails can better accommodate outliers. The result in Proposition 6 of [9] is extended from Gaussian to Student errors showing that the inverse regression approach of SIR is still valid outside the Gaussian case, meaning that the central subspace can still be estimated by maximum likelihood estimation of the inverse regression parameters. It is then shown that

the computation of the maximum likelihood estimators remains tractable in the Student case via an Expectation-Maximization (EM) algorithm which has a simple implementation and desirable properties.

The paper is organized as follows. In Section 2 general properties of the multivariate Student distribution and some of its variants are first recalled. The inverse regression model is introduced in Section 3 followed by the EM strategy to find the maximum likelihood estimator, the link with SIR and the resulting Student SIR algorithm. A simulation study is carried out in Section 4 and a real data application, showing the interest of this technique, is detailed in Section 5. The final section contains concluding remarks and perspectives. Proofs are postponed to the Appendix.

2. Multivariate generalized Student distributions

Multivariate Student, also called t-distributions, are useful when dealing with real-data because of their heavy tails. They are a robust alternative to the Gaussian distribution, which is known to be very sensitive to outliers. In contrast to the Gaussian case though, no closed-form solution exists for the maximum likelihood estimation of the parameters of the t-distribution. Tractability is, however, maintained both in the univariate and multivariate case, via the EM algorithm [17] and thanks to a useful representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture* [18]. A Gaussian scale mixture distribution has a probability density function of the form

$$P(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\psi}) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}/u)\ f_U(u;\boldsymbol{\psi})\ \mathrm{d}u, \tag{1}$$

where $\mathcal{N}_p(\ .\ ;\boldsymbol{\mu},\boldsymbol{\Sigma}/u)$ denotes the density function of the p-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/u$ and f_U is the probability distribution of a univariate positive variable U referred to hereafter as the weight variable. When f_U is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom, expression (1) leads to the standard p-dimensional t-distribution denoted by $t_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ with parameters $\boldsymbol{\mu}$ (lo-

cation vector), $\boldsymbol{\Sigma}$ ($p \times p$ positive definite scale matrix) and ν (positive degrees of freedom parameter). Its density is given by

$$\begin{aligned} t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) &= \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\ \mathcal{G}(u; \nu/2, \nu/2)\ \mathrm{d}u \\ &= \frac{\Gamma((\nu+p)/2)}{|\boldsymbol{\Sigma}|^{1/2}\ \Gamma(\nu/2)\ (\pi\nu)^{p/2}}\ [1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\nu]^{-(\nu+p)/2}, \end{aligned} \quad (2)$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. The Gamma distribution has probability density function $\mathcal{G}(u; \alpha, \gamma) = u^{\alpha-1}\Gamma(\alpha)^{-1} \exp(-\gamma u)\gamma^\alpha$, where Γ denotes the Gamma function.

If $f_U(u; \boldsymbol{\psi})$ is set equal to a Gamma distribution $\mathcal{G}(\alpha, \gamma)$ without imposing $\alpha = \gamma$, (1) results in a multivariate Pearson type VII distribution (see *e.g.* [19] vol.2 chap. 28) also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in [16]. This generalized version is the multivariate version of the t-distribution considered in this work, its density is given by:

$$\mathcal{S}_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \gamma) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\ \mathcal{G}(u; \alpha, \gamma)\ \mathrm{d}u \quad (3)$$

$$= \frac{\Gamma(\alpha + p/2)}{|\boldsymbol{\Sigma}|^{1/2}\ \Gamma(\alpha)\ (2\pi\gamma)^{p/2}}\ [1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/(2\gamma)]^{-(\alpha+p/2)}\ . \quad (4)$$

For a random variable $\mathbf{X}$ following distribution (4), an equivalent representation useful for simulation is $\mathbf{X} = \boldsymbol{\mu} + U^{-1/2}\tilde{\mathbf{X}}$ where U follows a $\mathcal{G}(\alpha, \gamma)$ distribution and $\tilde{\mathbf{X}}$ follows a $\mathcal{N}(0, \boldsymbol{\Sigma})$ distribution.

Remark 1 (Identifiability). *The expression* (4) *depends on* γ *and* $\boldsymbol{\Sigma}$ *only through the product* $\gamma\boldsymbol{\Sigma}$ *which means that to make the parameterization unique, an additional constraint is required. One possibility is to impose that* $\boldsymbol{\Sigma}$ *is of determinant 1. It is easy to see that this is equivalent to have an unconstrained* $\boldsymbol{\Sigma}$ *with* $\gamma = 1$.

Unconstrained parameters are easier to deal with in inference algorithms. Therefore, we will rather assume without loss of generality that $\gamma = 1$ with the notation $\mathcal{S}_p(0, \mathbf{V}, \alpha, 1) \equiv \mathcal{S}_p(0, \mathbf{V}, \alpha)$ adopted in the next Section.

3. Student Sliced Inverse Regression

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector, $Y \in \mathbb{R}$ the real response variable and $S_{Y|X}$ the central subspace spanned by the columns of the matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. In the following, it is assumed that $\dim(S_{Y|X}) = d$ where d is known and $d \leq p$. To address the estimation of the central subspace, we consider the inverse regression formulation of [9], which models the link from Y to $\mathbf{X}$. In addition to be a simpler regression problem, the inverse regression approach is of great interest because Proposition 6 in [9] states that in the Gaussian case, an estimation of the central subspace is provided by the estimation of the inverse regression parameters. In Subsection 3.1, the inverse regression model of [9] is extended by considering Student distributed errors. It is then shown in Subsection 3.2 that the estimation of the extended model is tractable via an Expectation-Maximization algorithm (EM). A link with SIR is presented in Subsection 3.3 and the resulting Student SIR algorithm is described in Subsection 3.4.

3.1. Student multi-index inverse regression model

In the spirit of [9, 10] the following regression model is considered

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{VBc}(Y) + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a non random vector, $\mathbf{B}$ is a non random $p \times d$ matrix with $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$, $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is a centered generalized Student random vector following the distribution given in (4), $\boldsymbol{\varepsilon}$ is assumed independent of Y, with scale matrix $\mathbf{V}$, $\mathbf{c} : \mathbb{R} \to \mathbb{R}^d$ is a non random function. It directly follows from (5) that

$$\mathbb{E}(\mathbf{X}|Y = y) = \boldsymbol{\mu} + \mathbf{VBc}(y), \tag{6}$$

and thus, after translation by $\boldsymbol{\mu}$, the conditional expectation of $\mathbf{X}$ given Y is a random vector located in the space spanned by the columns of $\mathbf{VB}$. When $\boldsymbol{\varepsilon}$ is assumed to be Gaussian distributed, Proposition 6 in [9] states that $\mathbf{B}$ corresponds to the directions of the central subspace $\boldsymbol{\beta}$. In [9, 10], it appears then that, under appropriate conditions, the maximum likelihood

estimator of $\mathbf{B}$ is (up to a full rank linear transformation) the SIR estimator of $\boldsymbol{\beta}$, *i.e.* Span$\{\mathbf{B}\}$ = Span$\{\boldsymbol{\beta}\}$. Proposition 6 in [9] can be generalized to our Student setting, so that $\mathbf{B}$ still corresponds to the central subspace. The generalization of Proposition 6 of [9] is given below.

Proposition 1. *Let $\mathbf{X}_y$ be a random variable distributed as $\mathbf{X}|Y = y$, let us assume that*

$$\mathbf{X}_y = \boldsymbol{\mu} + \mathbf{VBc}(y) + \boldsymbol{\varepsilon}, \tag{7}$$

with $\boldsymbol{\varepsilon}$ following a generalized Student distribution $\mathcal{S}_p(0, \mathbf{V}, \alpha)$, $\mathbf{c}(y) \in \mathbb{R}^d$ is function of y and $\mathbf{VB}$ is a $p \times d$ matrix of rank d. Under model (7), the distribution of $Y|\mathbf{X} = \mathbf{x}$ is the same as the distribution of $Y|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ for all values $\mathbf{x}$.

The proof is given in Appendix 7.1. According to this proposition, $\mathbf{X}$ can be replaced by $\mathbf{B}^T\mathbf{X}$ without loss of information on the regression of Y on $\mathbf{X}$. A procedure to estimate $\mathbf{B}$ is then proposed in the next Section

3.2. Maximum likelihood estimation via EM algorithm

Let $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$ be a set of independent random variables distributed according to the distribution of $(\mathbf{X}, Y)$ as defined in (5). The unknown quantities to be estimated in model (5) are $\{\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \alpha\}$ and the function $\mathbf{c}(.)$. Regarding $\mathbf{c}$, we focus on projection estimators for each coordinate of $\mathbf{c}(.) = (c_1(.), \ldots, c_d(.))$. For $k = 1, \ldots, d$, function $c_k(.)$ is expanded as a linear combination of h basis functions $s_j(.)$, $j = 1, \ldots, h$ as

$$c_k(.) = \sum_{j=1}^{h} c_{jk} s_j(.), \tag{8}$$

where the coefficients c_{jk}, $j = 1, \ldots, h$ and $k = 1, \ldots, d$ are unknown and to be estimated while h is supposed to be known. Let $\mathbf{C}$ be a $h \times d$ matrix with the kth column given by $(c_{1k}, \ldots, c_{hk})^T$ and $\mathbf{s}(.) = (s_1(.), \ldots, s_h(.))^T$. Then, model (5) can be rewritten as

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}(Y) + \boldsymbol{\varepsilon}, \qquad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{S}_p(0, \mathbf{V}, \alpha), \tag{9}$$

where $\mathcal{S}_p(0, \mathbf{V}, \alpha)$ is the multivariate centered generalized Student distribution with scale matrix $\mathbf{V}$. For each i, it follows that conditionally to Y_i, $\mathbf{X}_i \sim \mathcal{S}_p(\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}, \alpha)$ where $\mathbf{s}_i$=$\mathbf{s}(Y_i)$. The density of the generalized Student distribution is available in closed form and given in (4). However to perform the estimation, a more useful representation of this distribution is given by its Gaussian scale mixture representation (3). Introducing an additional set of latent variables $\mathbf{U} = \{U_1, \ldots, U_n\}$ with U_i independent of Y_i, one can equivalently write:

$$\mathbf{X}_i | U_i = u_i, Y_i = y_i \sim \mathcal{N}_p(\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}/u_i), \tag{10}$$

$$U_i | Y_i = y_i \sim \mathcal{G}(\alpha, 1). \tag{11}$$

Let us denote by $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}, \alpha\}$ the parameters to estimate from realizations $\{\mathbf{x}_i, y_i, i = 1, \ldots, n\}$. In contrast to the Gaussian case, the maximum likelihood estimates are not available in closed-form for the t-distributions. However, they are reachable using an Expectation-Maximization (EM) algorithm. More specifically, at iteration (t) of the algorithm, $\boldsymbol{\theta}$ is updated from a current value $\boldsymbol{\theta}^{(t-1)}$ to a new value $\boldsymbol{\theta}^{(t)}$ defined as $\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$. Considering the scale mixture representation above, a natural choice for Q is the following expected value of the complete log-likelihood:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &= E_U[\sum_{i=1}^{n} \log P(\mathbf{x}_i, U_i | Y_i = y_i; \boldsymbol{\theta}) | \mathbf{X}_i = \mathbf{x}_i, Y_i = y_i; \boldsymbol{\theta}^{(t-1)}] \qquad (12)\\
&= \sum_{i=1}^{n} E_{U_i}[\log P(\mathbf{x}_i | U_i, y_i; \boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] + E_{U_i}[\log P(U_i; \alpha) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}]\\
&= -\frac{1}{2} n \log\det \mathbf{V} + \frac{1}{2} p \sum_{i=1}^{n} E_{U_i}[\log(U_i) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}]\\
&- \frac{1}{2} \sum_{i=1}^{n} E_{U_i}[U_i | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \; (\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i - \mathbf{x}_i)^T \mathbf{V}^{-1} (\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i - \mathbf{x}_i)\\
&+ \sum_{i=1}^{n} E_{U_i}[\log P(U_i; \alpha) | \mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \; .
\end{aligned}$$

Note that all computations are conditionally to the Y_i's and no assumption

is made on the distribution of the Y_i's. The E-step therefore consists of computing the quantities

$$\bar{u_i}^{(t)} = E_{U_i}[U_i|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \,, \tag{13}$$

$$\tilde{u_i}^{(t)} = E_{U_i}[\log U_i|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \,, \tag{14}$$

while the M-step divides into two-independent M-steps involving separately parameters $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ and α. The second quantity (14) is needed only in the estimation of α. The following notation is introduced for the next sections:

$$\bar{u}^{(t)} = \frac{\sum_{i=1}^n \bar{u_i}^{(t)}}{n} \tag{15}$$

$$\tilde{u}^{(t)} = \frac{\sum_{i=1}^n \tilde{u_i}^{(t)}}{n} \,. \tag{16}$$

E-step. The quantities (13) and (14) above require the posterior distribution of the U_i's. This distribution can be easily determined using the well known conjugacy of the Gamma and Gaussian distributions for the mean. It follows then from standard Bayesian computations that the posterior distribution is still a Gamma distribution with parameters specified below,

$$\begin{aligned}
&p(u_i|\mathbf{X}_i = \mathbf{x}_i, Y_i = y_i; \boldsymbol{\theta}^{(t-1)}) \\
&\quad \propto \mathcal{N}_p(\mathbf{x}_i; \boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)}/u_i)\ \mathcal{G}(u_i; \alpha^{(t-1)}, 1) \\
&\quad = \mathcal{G}(u_i; \alpha^{(t-1)} + \frac{p}{2}, 1 + \frac{1}{2}\delta(\mathbf{x}_i, \mu^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})),
\end{aligned}$$

where $\delta(\mathbf{x}_i, \boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i, \mathbf{V}) = (\boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i-\mathbf{x}_i)^T\mathbf{V}^{-1}(\boldsymbol{\mu}+\mathbf{VBC}^T\mathbf{s}_i-\mathbf{x}_i)$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\boldsymbol{\mu} + \mathbf{VBC}^T\mathbf{s}_i$ when the covariance is $\mathbf{V}$.

The required moments (13) and (14) are then well known for a Gamma distribution, so that it comes,

$$\bar{u}_i^{(t)} = \frac{\alpha^{(t-1)} + \frac{p}{2}}{1 + \frac{1}{2}\delta(\mathbf{x}_i, \mu^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})} \text{ and}$$

$$\tilde{u}_i^{(t)} = \Psi(\alpha^{(t-1)} + \frac{p}{2}) - \log(1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i, \mathbf{V}^{(t-1)})) \,,$$

where Ψ is the Digamma function. As it will become clear in the following M-step, $\bar{u}_i^{(t)}$ acts as a weight for $\mathbf{x}_i$. Whenever the Mahalanobis distance of $\mathbf{x}_i$ to $\boldsymbol{\mu}^{(t-1)} + \mathbf{V}^{(t-1)}\mathbf{B}^{(t-1)}\mathbf{C}^{(t-1)T}\mathbf{s}_i$ increases, the weight $\bar{u}_i^{(t)}$ of $\mathbf{x}_i$ decreases and the influence of $\mathbf{x}_i$ in the estimation of the parameters will be downweighted in the next iteration. The idea of using weights to handle outliers is common in the literature, Weighted Inverse Regression (WIRE) [15] gives weights through a deterministic kernel function to ensure the existence of the first moment. Our approach does not require previous knowledge to select an appropriate kernel and refers to the wide range of t-distributions (the Cauchy distribution for which the first moment is not defined lies in this family).

M- step. The M-step divides into the following two independent sub-steps. **M-$(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ substep.** Omitting terms that do not depend on the parameters in (12), estimating $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ by maximization of Q consists, at iteration (t), of minimizing with respect to $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ the following G function,

$$G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) = \log\det \mathbf{V} + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}^{(t)} \, (\boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i)^T\mathbf{V}^{-1}(\boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i) \, . \quad (17)$$

To this aim, let us introduce (omitting the index iteration (t) in the notation) the $h \times h$ weighted covariance matrix $\mathbf{W}$ of $\mathbf{s}(Y)$ defined by:

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \, (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T,$$

the $h \times p$ weighted covariance matrix $\mathbf{M}$ of $(\mathbf{s}, \mathbf{X})$ defined by

$$\mathbf{M} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \, (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

and $\boldsymbol{\Sigma}$ the $p \times p$ weighted covariance matrix of $\mathbf{X}$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i} \, (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \quad (18)$$

where

$$\bar{\mathbf{x}} = \frac{1}{\sum_{i=1}^{n} \bar{u}_i} \sum_{i=1}^{n} \bar{u}_i \mathbf{x}_i \quad \text{and} \tag{19}$$

$$\bar{\mathbf{s}} = \frac{1}{\sum_{i=1}^{n} \bar{u}_i} \sum_{i=1}^{n} \bar{u}_i \mathbf{s}_i. \tag{20}$$

We derive then the following lemma.

Lemma 1. *Using the above notations,* $G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ *can be rewritten as*

$$G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) = \log\det \mathbf{V} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{V}^{-1}) + \mathrm{tr}(\mathbf{C}^T\mathbf{W}\mathbf{C}\mathbf{B}^T\mathbf{V}\mathbf{B}) - 2\mathrm{tr}(\mathbf{C}^T\mathbf{M}\mathbf{B}) + \bar{u}\ (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})^T\mathbf{V}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})\ .$$

The proof is given in Appendix 7.2. Thanks to this representation of $G(.)$ it is possible to derive the following proposition which is a generalization to the multi-index case and Student setting of the result obtained in case of Gaussian error ϵ in [10].

Proposition 2. *Under (9), if* $\mathbf{W}$ *and* $\boldsymbol{\Sigma}$ *are regular, then the M-step for* $(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ *leads to the updated estimations* $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ *given below*

- $\hat{\mathbf{B}}$ *is made of the eigenvectors associated to the largest eigenvalues of* $\boldsymbol{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$,
- $\hat{\mathbf{V}} = \boldsymbol{\Sigma} - (\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})(\mathbf{B}^T\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})^{-1}(\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\mathbf{B})^T$,
- $\hat{\mathbf{C}} = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}$ *and*
- $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}$.

The proof is detailed in Appendix 7.3. Regarding parameter α it can be updated using an independent part of Q as detailed in the next M-step.

M-α substep.

Parameter α can be estimated by maximizing independently with regards to α,

$$\sum_{i=1}^{n} E_{U_i}[\log P(U_i;\alpha)|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] \ . \tag{21}$$

Then, since

$$E_{U_i}[\log p(U_i;\alpha)|\mathbf{x}_i, y_i; \boldsymbol{\theta}^{(t-1)}] = -\bar{u_i}^{(t)} + (\alpha-1)\tilde{u_i}^{(t)} - \log\Gamma(\alpha) \ , \tag{22}$$

setting the derivative with respect to α to zero, we obtain that $\hat{\alpha} = \Psi^{-1}(\tilde{u})$, where $\Psi(.)$ is the Digamma function.

In practice, for the procedure to be complete, the choice of the h basis functions s_j needs to be specified. Many possibilities for basis functions are available in the literature such as classical Fourier series, polynomials, etc. In the next section, we discuss a choice of basis functions which provides the connection with Sliced Inverse Regression (SIR) [3].

3.3. Connection to Sliced Inverse Regression

As in the Gaussian case [9, 10], a clear connection with SIR can be established for a specific choice of the h basis functions. When Y is univariate a natural approach is to first partition the range of Y into $h+1$ bins S_j for $j = 1, \ldots, h+1$ also referred to as slices, and then defining h basis functions by considering the first h slices as follows,

$$s_j(.) = \mathbb{1}\{. \in S_j\}, \quad j = 1, \ldots, h, \tag{23}$$

where $\mathbb{1}$ is the indicator function. Note that it is important to remove one of the slices so that the basis functions remain independent. However, the following related quantities are defined for $j = 1, \ldots, h+1$:

$$\begin{aligned} n_j &= \sum_{i=1}^{n} \bar{u}_i \mathbb{1}\{y_i \in S_j\}, \\ f_j &= \frac{n_j}{n} \ . \end{aligned} \tag{24}$$

They represent respectively the number of y_i in slice j weighted by the $\bar{u}_i$ and the weighted proportion in slice j. The following weighted mean of $\mathbf{X}$ given $Y \in S_j$ is then denoted by

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n} \bar{u}_i \mathbb{1}\{y_i \in S_j\} \mathbf{x}_i, \tag{25}$$

and the $p \times p$ "between slices" covariance matrix by

$$\mathbf{\Gamma} = \sum_{j=1}^{h+1} f_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T.$$

In this context, the following consequence of Proposition 2 can be established.

Corollary 1. *Under (9) and (23), if* $\mathbf{\Sigma}$ *is regular, then the updated estimation* $\hat{\mathbf{B}}$ *of* $\mathbf{B}$ *is given by the eigenvectors associated to the largest eigenvalues of* $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}$. *In addition,* $\mathbf{\Gamma} = \mathbf{M}^T \mathbf{W}^{-1} \mathbf{M}$.

The proof is given in Appendix 7.4. When all $\bar{u}_i = 1$, the iterative EM algorithm reduces to one M-step and the quantities defined in this section correspond to the standard SIR estimators. The EM algorithm resulting from this choice of basis functions is referred to as the Student SIR algorithm. It is outlined in the next section.

3.4. Central subspace estimation via Student SIR algorithm

The EM algorithm can be outlined using Proposition 2 and Corollary 1. It relies on two additional features to be specified, initialization and stopping rule. As the algorithm alternates the E and M steps, it is equivalent to start with one of this step. It is convenient to start with the Maximization step since the initialization of quantities $\bar{u}_i, \tilde{u}_i$ can be better interpreted. If $\bar{u}_i$ is constant and $\tilde{u}_i = 0$, the first M-step of the algorithm results in performing standard SIR. Regarding an appropriate stopping rule of the algorithm, EM's fundamental property is to increase the log-likelihood at each iteration. A standard criteria is then the relative increase in log-likelihood, denoted by

$\Delta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)})$, between two iterations. At each iteration, for current parameters values, the log-likelihood is easy to compute using (4) and (9). Another natural criterion is to assess when parameter estimation stabilizes. Typically, focusing on the central subspace $\mathbf{B}$, the following proximity measure [20, 21] can be considered:

$$r(\mathbf{B}, \hat{\mathbf{B}}) = \frac{trace(\mathbf{B}\mathbf{B}^T\hat{\mathbf{B}}\hat{\mathbf{B}}^T)}{d}. \tag{26}$$

The above quantity r ranges from 0 to 1 and evaluates the distance between the subspaces spanned by the columns of $\mathbf{B}$ and $\hat{\mathbf{B}}$. If $d = 1$, r is the squared cosine between the two spanning vectors. Although not directly related to the EM algorithm, in practice this criterion gave similar results in terms of parameter estimation. Experiments on simulated and real data are reported in the next two sections.

3.5. Determination of the central subspace dimension

Determining the dimension d of the central subspace is an important issue for which different solutions have been proposed in the literature. Most users rely on graphical considerations, *e.g.* [22]. A more quantitative approach is to use cross validation after the link function is found. Although in that case, d may vary depending on the specific regression approach that the user selected. Other methods that can be easily used on real data, are mainly based on (sequential) tests [3, 23, 24, 20, 25, 11]. An alternative that uses a penalized likelihood criterion has been proposed in [26]. In our setting, formulated as a maximum likelihood problem, the penalized likelihood approach is the most natural. For a given value d of the central subspace dimension, we therefore propose to compute the Bayesian information criterion [27] defined as $BIC(d) = -2L(d) + \eta \log n$, where $\eta = \frac{p(p+3)}{2} + 1 + \frac{d(2p-d-1+2h)}{2}$ is the number of free parameters in the model and $L(d)$ is the maximized log-likelihood computed at the parameters values obtained at convergence of the EM algorithm. Computing $L(d)$ is a straightforward byproduct of the algorithm described above as this quantity is already used in our stopping

Algorithm 1 Student SIR algorithm

Set h and partition the Y range into $h+1$ slices.

Set the e.d.r. space dimension d and the desired tolerance value for convergence δ.

Initialize the $\bar{u}_i^{(0)}, \tilde{u}_i^{(0)}$'s with $\bar{u}_i^{(0)} = 1$ and $\tilde{u}_i^{(0)} = 0$ for all $i = 1, \ldots, n$

(this first iteration of the algorithm gives the SIR estimation of Γ and $\mathbf{B}$).

while $\Delta(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)}) < \delta$ **do**

M-step

Compute:

- $\bar{u}^{(t)}$ and $\tilde{u}^{(t)}$ (eq. (15) and (16)), $f^{(t)} = (f_1^{(t)}, \ldots f_h^{(t)})^T$ and $f_{h+1}^{(t)}$ (eq. (24)),
- $\bar{\mathbf{x}}_j^{(t)}$ and $\bar{\mathbf{x}}^{(t)}$ (eq. (25) and (19)),
- $\boldsymbol{\Sigma}^{(t)}$ (eq. (18)),
- $\mathbf{M}^{(t)}$ where each row is given by $\mathbf{M}_{j,\cdot}^{(t)} = f_j^{(t)}(\bar{\mathbf{x}}_j^{(t)} - \bar{\mathbf{x}}^{(t)})^T$ for $j = 1, \ldots, h$,
- $\mathbf{W}^{(t)-1} = \operatorname{diag}\left(\frac{1}{f_1^{(t)}}, \ldots, \frac{1}{f_h^{(t)}}\right) + \frac{1}{f_{h+1}^{(t)}}\mathbf{O}$, where $\mathbf{O}$ is the $h \times h$ matrix defined by $O_{ij} = 1$,
- $\boldsymbol{\Gamma}^{(t)} = \mathbf{M}^{(t)T}\mathbf{W}^{(t)-1}\mathbf{M}^{(t)}$,
- $\mathbf{B}^{(t)}$ matrix of the d eigenvectors associated to the d largest eigenvalues of $\boldsymbol{\Sigma}^{(t)-1}\boldsymbol{\Gamma}^{(t)}$,
- $\mathbf{V}^{(t)} = \boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)}(\mathbf{B}^{(t)^T}\boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)})^{-1}(\boldsymbol{\Gamma}^{(t)}\mathbf{B}^{(t)})^T$,
- $\mathbf{C}^{(t)} = \mathbf{W}^{(t)-1}\mathbf{M}^{(t)}\mathbf{B}^{(t)}(\mathbf{B}^{(t)T}\mathbf{V}^{(t)}\mathbf{B}^{(t)})^{-1}$,
- $\boldsymbol{\mu}^{(t)} = \bar{\mathbf{x}}^{(t)} - \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\bar{\mathbf{s}}^{(t)}$,
- $\alpha^{(t)} = \Psi^{-1}(\tilde{u}^{(t)})$.

E-step

Update the $\bar{u}_i, \tilde{u}_i$'s using the quantities estimated in the M-step:

$$\bar{u_i}^{(t+1)} = \frac{\alpha^{(t)} + \frac{p}{2}}{1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t)} + \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\mathbf{s}_i, \mathbf{V}^{(t)})},$$

$$\tilde{u_i}^{(t+1)} = \Psi(\alpha^{(t)} + \frac{p}{2}) - \log(1 + \frac{1}{2}\delta(\mathbf{x}_i, \boldsymbol{\mu}^{(t)} + \mathbf{V}^{(t)}\mathbf{B}^{(t)}\mathbf{C}^{(t)T}\mathbf{s}_i, \mathbf{V}^{(t)})).$$

end while

criterion. Following the BIC principle, an estimator of d can then be defined as the minimizer of $BIC(d)$ over $d \in \{1, \ldots, min(p, h)\}$. The performance of this criterion is investigated in the simulation study in Section 4 and used on the real data example of Section 5. The simulation study reveals that BIC can provide correct selections but requires large enough sample sizes. This limitation has been already pointed out in the literature (see *e.g.* [28]).

4. Simulation study

Student SIR is tested on simulated data under a variety of different models and distributions for the p-dimensional random variable $\mathbf{X}$. The behavior of Student SIR is compared to SIR and four other techniques arising from the literature that claim some robustness. For comparison, the simulation setup described in [15, 14] is adopted.

4.1. Simulation setup

Three different regression models are considered:

I : $Y = 1 + 0.6X_1 - 0.4X_2 + 0.8X_3 + 0.2\varepsilon$,

II : $Y = (1 + 0.1\varepsilon)X_1$,

III : $Y = X_1/(0.5 + (X_2 + 1.5)^2)) + 0.2\varepsilon$,

where ε follows a standard normal distribution. The three models are combined with three possible distributions for the predictors $\mathbf{X}$:

(i) $\mathbf{X}$ is multivariate normal distributed with mean vector $\mathbf{0}$ and covariance matrix defined by its entries as $\sigma_{ij} = 0.5^{|i-j|}$;

(ii) $\mathbf{X}$ is standard multivariate Cauchy distributed;

(iii) $\mathbf{X} = (X_1, \ldots, X_p)^T$, where each X_i is generated independently from a mixture of normal and uniform distributions denoted by $0.8\mathcal{N}(0, 1) + 0.2\mathcal{U}(-\nu, \nu)$ where ν is a positive scalar value.

Models **I**, **III** are homoscedastic while model **II** is heteroscedastic. Case (**ii**) is built to test the sensitivity to outliers while the distribution of $\mathbf{X}$ is elliptical. In (**iii**) a non-elliptical distribution of $\mathbf{X}$ is considered. The dimension is set to $p = 10$, the dimension of the e.d.r. space is $d = 1$ for **I**, **II** and $d = 2$ for **III**. The nine different configurations of $\mathbf{X}$ and Y are simulated with a number of samples varying depending on the experiment. In all tables Student SIR is compared with standard SIR and four other approaches. Contour Projection (CP-SIR) [29, 30] applies the SIR procedure on a rescaled version of the predictors. Weighted Canonical Correlation (WCAN) [31] uses a basis of B-splines first estimating the dimension d of the central subspace and then the directions from the nonzero robustified version of the correlation matrices between the predictors and the B-splines basis functions. The idea of Weighted Inverse Regression (WIRE) [15] is to use a different weight function capable of dealing with both outliers and inliers. SIR is a particular case of WIRE with constant weighting function. Slice Inverse Median Estimation (SIME) replaces the intra slice mean estimator with the median which is well known to be more robust. All values referring to CP-SIR, WCAN, WIRE, SIME in the tables are directly extracted from [15]. Values relative to SIR have been recomputed using [32].

4.2. Results

To assess the sensitivity of the compared methods to different setting parameters, four sets of tests are carried out and reported respectively in Tables 1 and 2. First, the 9 configurations of $\mathbf{X}$ and Y models are tested for fixed sample size $n = 200$, number of slices $h = 5$ and $p = 10$ (Table 1 (a)). Then, the effect of the sample size is illustrated for model **I** (Table 1 (b)). The number of slices is varied to evaluate the sensitivity to the h value (Table reftb3 (a)) and at last, different values of ν are tested in the model (**iii**) case (Table 2 (b)). In all cases and tables, the different methods performance is assessed based on their ability to recover the central subspace which is measured via the value of the proximity measure r (26).

Student SIR shows its capability to deal with different configurations.

The proximity criterion (26) in Table 1 (a) is very close to one, for the first two regression models independently of the distributions of the predictors. In the Gaussian case, Student SIR and SIR are performing equally well showing that our approach has no undesirable effects when dealing with *simple* cases. For configuration $\mathbf{III}-(\mathbf{iii})$, a slightly different value has been found for SIR compared to [15]. In this configuration however the trend is clear: standard SIR, Student SIR, WIRE and SIME show similar performance. In contrast, configurations $\mathbf{I}-(\mathbf{ii}), \mathbf{II}-(\mathbf{ii}), \mathbf{III}-(\mathbf{ii})$ illustrate that Student SIR can significantly outperform SIR. This is not surprising since the standard multivariate Cauchy has heavy tails and SIR is sensitive to outliers [14].

Table 1 (b) illustrates on model $\mathbf{I}$ the effect of the sample size n: Student SIR exhibits the best performance among all methods. It is interesting to observe that, in case (**ii**), the smaller value of r for standard SIR does not depend on the sample size n. In contrast, adding observations results in a better estimation for Student SIR.

It is then known that SIR is not very sensitive to the number of slices h [22]. In Table 2 (a), an analysis is performed with varying h. Student SIR appears to be as well not very sensitive to the number of slices.

Extra inliers as well as outliers can affect the estimation. In case (**iii**), parameter ν is controlling the extra observations magnitude. Under different values of $\nu = 0.5, 0.2, 0.1, 0.05$,Table 2 (b) shows that both SIR and Student SIR are robust to inliers while CP-SIR and WCAN fail when ν is small and extra observations behave as inliers concentrated around the average.

In addition, a study on the behavior of SIR and Student SIR when $\mathbf{X}$ follows a standard multivariate Student distribution, with different degrees of freedom (df), is shown in Table 3 (a). The multivariate Cauchy of model (**ii**) coincides with the multivariate Student with one degree of freedom. This setting is favorable to our model which is designed to handle heavy tails. Not surprisingly, Student SIR provides better results for small degrees of freedom but the difference with SIR is reduced as the degree of freedom increases and the multivariate Student gets closer to a Gaussian. The stan-

dard deviation follows the same trend. In case $\mathbf{III}-(\mathbf{ii})$ the convergence of SIR becomes extremely slow. Regarding computational time, results are reported in Table 3 (b). Student SIR has multiple iterations, which increases computational time compared to SIR. It is interesting that, in the cases in which SIR fails $(\mathbf{I}-(\mathbf{ii}), \mathbf{II}-(\mathbf{ii}), \mathbf{III}-(\mathbf{ii})$ see Table 1), the convergence of Student SIR is fast, requiring less than a second on a standard laptop (Our Matlab code is available at https://hal.inria.fr/hal-01294982). All reported results have been obtained using a threshold of 0.01 for the relative increase of the Log-likelihood.

At last, the use of BIC as a selection criterion for the central subspace dimension d is investigated. As an illustration, last column of Table 3 (b) shows the number of times the criterion succeeded in selecting the correct dimension (*i.e.* $d = 2$ in this example) over 200 repetitions. BIC performs very well provided the sample size is large enough, this phenomenon being more critical as the number of outlying data increases. This is not surprising as this limitation of BIC has often been reported in the literature.

To summarize, through these simulations Student SIR shows good performance, outperforming SIR when the distribution of $\mathbf{X}$ is heavy-tailed (case $(\mathbf{ii})$) and preserving good properties such as insensitivity to the number of slices or robustness to inliers that are peculiar of SIR.

5. Real data application: The galaxy dataset

5.1. Data

The Galaxy dataset corresponds to $n = 362,887$ different galaxies. This dataset has been already used in [33] with a preprocessing based on expert supervision to remove outliers. In this study all the original observations are considered, removing only points with missing values, which requires no expertise. The response variable Y is the stellar formation rate. The predictor $\mathbf{X}$ is made of spectral characteristics of the galaxies and is of dimension $p = 46$.

5.2. Evaluation setting

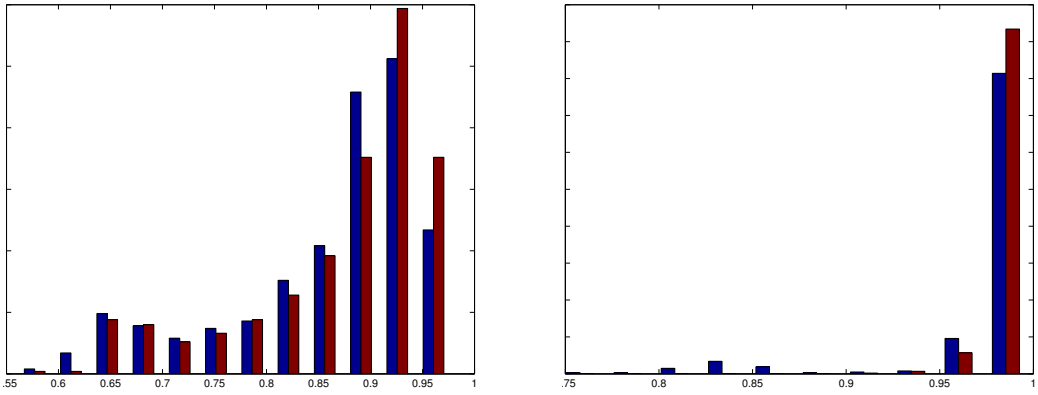
The number of samples n is very large and the proportion of outliers is very small compared to the whole dataset. The following strategy is adopted: 1000 random subsets of $\mathbf{X}$ of size $n_a = 3,000$, $\mathbf{X}_i^a$, $i = 1, \ldots, 1000$ and size $n_b = 30,000$, $\mathbf{X}_i^b$, $i = 1, \ldots, 1000$ are considered to compare the performance of SIR and Student SIR. First a reference result $\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}^{\text{st-SIR}}$ is obtained using the whole dataset $\mathbf{X}$, using respectively SIR and Student SIR with the dimension of the e.d.r. space set to $d = 3$ and the number of slices to $h = 1000$. The value $d = 3$ was selected via BIC computed for $d = 1, ..., 20$, which is reliable for such a large sample size. The proximity measure r (26) between the two reference spaces is $r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}^{\text{st-SIR}}) = 0.95$. SIR and st-SIR are identifying approximately the same same e.d.r. space.

5.3. Results

Let $\hat{\mathbf{B}}_i^{\text{SIR}}$, $\hat{\mathbf{B}}_i^{\text{st-SIR}}$ be the estimations of the basis of the e.d.r. space for the random subsets $\mathbf{X}_i^a$, $i = 1, \ldots, 1000$ using respectively SIR and Student SIR. The proximity measures $r_i^{\text{SIR}} = r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}_i^{\text{SIR}})$ and $r_i^{\text{st-SIR}} = r(\hat{\mathbf{B}}^{\text{st-SIR}}, \hat{\mathbf{B}}_i^{\text{st-SIR}})$ are considered. All results are obtained setting the number of slices to $h = 10$. The means (and standard deviations) of the resulting proximity measures r are respectively 0.86(0.09) for SIR and 0.87(0.09) for Student SIR. The experiment is better visualized in Figure 1 (a) where histograms show that Student SIR performs better than SIR most of the time. As expected SIR is less robust than Student SIR, obtaining with a higher frequency low values of r. The histograms show a difference between values around $r = 0.96$ (23.8% of random subsets for Student SIR, 17.2% for SIR).

In the second test, the sample size of the subsets is increased to $n_b = 30,000$. Accordingly, the number of slices is increased to $h = 100$. Not surprisingly, the means (and standard deviations) of r_i^{SIR} and $r_i^{\text{st-SIR}}$ are increasing to 0.97(0.04) and 0.99(0.00). Student SIR however still performs better than SIR (Figure 1 (b)) with some low values of the proximity measure for SIR while Student SIR has almost all the values (93.4% of random

subsets) concentrated around $r = 0.98$. The difference between the two approaches is then further emphasized in Figure 2 where the cloud of points in the upper left corner of the plot corresponds to datasets for which SIR was not able to estimate a correct basis of the e.d.r space while Student SIR shows good performance. Even if the true e.d.r space is unknown, this analysis suggests that Student SIR is robust to outliers and can be profitably used in real applications.



(a) (b)

Figure 1: Histograms of the proximity measure (26) $r_i^{\text{SIR}} = r(\hat{\mathbf{B}}^{\text{SIR}}, \hat{\mathbf{B}}_i^{\text{SIR}})$ (blue) and $r_i^{\text{st-SIR}} = r(\hat{\mathbf{B}}^{\text{st-SIR}}, \hat{\mathbf{B}}_i^{\text{st-SIR}})$ (red) for $i = 1, \ldots, 1000$ random subsets of $\mathbf{X}$ of size n_a=3000 **(a)** and n_b=30,000 **(b)**.

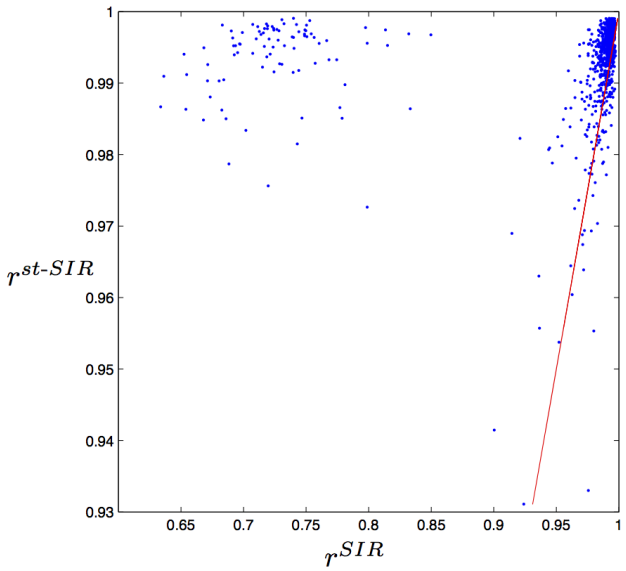


Figure 2: Horizontal axis r_i^{SIR}, vertical axis $r_i^{\text{st-SIR}}$, $i = 1, \ldots, 1000$, proximity measures computed using subsets of $\mathbf{X}$ of size $n_b = 30,000$. Almost all points are lying above the line $y = x$ indicating that Student SIR improves SIR results and significantly so for the subsets in the upper left corner.

Model	**X**	Method					
		SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	**.99(.01)**	.99(.01)	.98(.01)	.98(.01)	**.99(.01)**	**.99(.01)**
	(ii)	.63(.18)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
	(iii)	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.01)**
II	(i)	**.99(.01)**	.98(.01)	.98(.01)	.98(.01)	98(.01)	**.99(.01)**
	(ii)	.61(.18)	.92(.04)	.89(.06)	.87(.08)	.91(.05)	**.98(.01)**
	(iii)	**.99(.01)**	.67(.25)	.69(.28)	.98(.01)	.97(.02)	**.99(.01)**
III	(i)	.88(.06)	.87(.06)	**.89(.05)**	.86(.06)	.87(.06)	.87(.06)
	(ii)	.40(.13)	.78(.10)	.78(.11)	.76(.11)	.78(.10)	**.85(.06)**
	(iii)	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)

(a)

Model	**X**	n	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	50	**.95(.03)**	.91(.09)	.86(.11)	.88(.11)	.90(.08)	**.95(.03)**
		100	**.98(.01)**	.96(.03)	.96(.03)	.95(.03)	.96(.02)	**.98(.01)**
		200	**.99(.01)**	**.99(.01)**	.98(.01)	.98 (.01)	**.99(.01)**	**.99(.01)**
		400	**1(.00)**	.99(.00)	.99(.00)	.99 (.01)	.99(.00)	**1(.00)**
	(ii)	50	.60(.22)	.66(.18)	.57(.23)	.49(.24)	.59(.21)	**.90(.07)**
		100	.62(.21)	.85 (.08)	.78(.11)	.73(.15)	.81(.10)	**.96(.02)**
		200	.62(.20)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
		400	.62(.18)	.96(.02)	.94(.03)	.93(.03)	.96(.02)	**.99(.00)**
	(iii)	50	**.95(.02)**	.45(.29)	.18(.19)	.73(.25)	.86(.09)	**.95(.02)**
		100	**.98(.01)**	.66(.25)	.35(.29)	.94(.04)	.94(.04)	**.98(.01)**
		200	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.00)**
		400	**.99(.00)**	.96(.04)	.96(.04)	.93(.03)	**.99(.01)**	**.99(.00)**

(b)

Table 1: **(a)** Average of the proximity measure r (eq. (26)) for sample size $n = 200$; and **(b)** effect of sample size n on the average proximity measure r, both over 200 repetitions with standard deviation in brackets. Six methods are compared. SIR: sliced inverse regression; CP-SIR: contour projection for SIR; WCAN: weighted canonical correlation; WIRE: weighted sliced inverse regression estimation; SIME: sliced inverse multivariate median estimation and st-SIR: Student SIR. In all cases, the number of slices is $h = 5$ and the predictor dimension $p = 10$. Best r values are in bold.

Model	**X**	h	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	2	**.96(.02)**	.95(.03)	.98(.01)	.94(.03)	.95(.03)	.95(.02)
		5	**.99(.01)**	.98(.01)	.98(.01)	.98(.02)	.98(.01)	**.99(.00)**
		10	.99(.00)	.99(.01)	.98(.01)	.98 (.01)	.98(.01)	**1(.00)**
		20	**1(.00)**	.99(.01)	.98(.02)	.98 (.02)	.98(.01)	**1(.00)**
	(ii)	2	.60(.18)	.90(.05)	.60(.34)	.87(.06)	.89(.06)	**.95(.02)**
		5	.62(.18)	.92 (.04)	.89(.06)	.88(.07)	.92(.04)	**.98(.01)**
		10	.63(.19)	.92(.04)	.88(.07)	.87(.07)	.86(.08)	**.99(.00)**
		20	.65(.21)	.91(.05)	.85(.08)	.85(.08)	.69(.14)	**1(.00)**
	(iii)	2	**.96(.02)**	.91(.06)	.84(.20)	.95(.02)	.94(.05)	.95(.02)
		5	**.99(.00)**	.64(.26)	.67(.28)	.98(.01)	.98(.01)	**.99(.00)**
		10	**1(.00)**	.63(.26)	.48(.31)	.98(.01)	.98(.01)	**1(.00)**
		20	**1(.00)**	.53(.28)	.43(.30)	.98(.01)	.98(.01)	**1(.00)**

(a)

Model	Y	ν	Method					
			SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
(iii)	I	.5	**.99(.01)**	.98(.01)	.96(.02)	.96(.02)	.98(.01)	**.99(.01)**
		.2	**.99(.01)**	.96(.02)	.87(.15)	.97(.01)	.97(.01)	**.99(.01)**
		.1	**.99(.01)**	.86(.12)	.72(.27)	.98 (.01)	.97(.01)	**.99(.01)**
		.05	**.99(.01)**	.58(.24)	.65(.30)	.98 (.01)	.97(.01)	**.99(.01)**
	II	.5	**.99(.01)**	.98(.01)	.96(.02)	.96(.02)	.98(.01)	**.99(.01)**
		.2	**.99(.01)**	.96 .03)	.86(.16)	.98(.01)	.98(.01)	**.99(.01)**
		.1	**.99(.01)**	.67(.25)	.69(.28)	98(.01)	.97(.02)	**.99(.01)**
		.05	**.99(.01)**	.28(.24)	.59(.29)	98(.01)	.97(.01)	**.99(.01)**
	III	.5	**.88(.06)**	.85(.07)	.84(.08)	.77(.11)	.87(.06)	**.88(.05)**
		.2	.84(.07)	.76(.12)	.71(.13)	.84(.08)	**.86(.06)**	.84(.07)
		.1	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)
		.05	.83(.07)	.58(.10)	.65(.13)	**.86(.07)**	**.86(.07)**	.82(.07)

(b)

Table 2: Effect of the number of slices **(a)** and of inlier magnitude ν **(b)** on the average proximity measure r (eq. (26)), over 200 repetitions with related standard deviation in brackets. Six methods are compared. SIR: sliced inverse regression; CP-SIR: contour projection for SIR; WCAN: weighted canonical correlation; WIRE: weighted sliced inverse regression estimation; SIME: sliced inverse multivariate median estimation and st-SIR: Student SIR. In all cases, the sample size is $n = 200$ and the predictor dimension $p = 10$. Best r values are in bold.

Model - **X**	*df*	Method	
		SIR	st-SIR
I - (ii)	3	.94(.05)	**.99(.00)**
	5	.98(.02)	**.99(.00)**
	7	.98(.01)	**.99(.00)**
	10	.99(.01)	**.99(.00)**
II - (ii)	3	.94(.05)	**.99(.00)**
	5	.97(.02)	**.99(.00)**
	7	.98(.01)	**.99(.00)**
	10	.99(.01)	**.99(.00)**
III - (ii)	3	.82(.08)	**.90(.04)**
	5	.88(.05)	**.92(.03)**
	7	.90(.04)	**.92(.03)**
	10	.90(.04)	**.92(.03)**
	30	.91(.03)	**.92(.03)**

(a)

Model - **X**	n	Method		
		SIR	st-SIR	BIC
III-(i)	200	.00(.00)	.13(.05)	25/200
	300	.01(.00)	.09(.03)	109/200
	400	.04(.01)	.33(.16)	156/200
	500	.05(.01)	.43(.17)	189/200
	1000	.10(.02)	.51(.17)	200/200
III-(ii)	200	.00(.00)	.13(.05)	21/200
	300	.01(.00)	.09(.05)	19/200
	400	.04(.01)	.33(.20)	39/200
	500	.05(.01)	.43(.18)	90/200
	1000	.10(.02)	.51(.20)	200/200
III-(iii)	200	.00(.00)	.13(.05)	0/200
	300	.01(.00)	.13(.04)	12/200
	400	.04(.01)	.38(.16)	22/200
	500	.05(.01)	.34(.13)	16/200
	1000	.10(.02)	.51(.17)	198/200

(b)

Table 3: **(a)** Effect of the degree of freedom (df) on the average of the proximity measure r (eq.(26)) for sample size $n = 200$, the number of slices is $h = 5$ and the predictor dimension $p = 10$; and **(b)** Effect of the sample size on the computational time in seconds (standard deviations in brackets) and ratio of correct selections ($d = 2$) for BIC over 200 runs.

6. Conclusion and future work

We proposed a new approach referred to as Student SIR to robustify SIR. In contrast to most existing approaches which aim at replacing the standard SIR estimators by robust versions, we considered the intrinsic characterization of SIR as a Gaussian inverse regression model [9] and modified it into a Student model with heavier tails. While SIR is not robust to outliers, Student SIR has shown to be able to deal with different kind of situations that depart from normality. As expected, when SIR provides good results, Student SIR is performing similarly but at a higher computational cost due to the need for an EM iterative algorithm for estimation.

Limitations of the approach include the difficulty in dealing with the case $p > n$ or when there are strong correlations between variables. Student SIR as well as SIR still suffer from the need to inverse large covariance matrices. A regularization, to overcome this problem, has been proposed in [10] and could be extended to our Student setting. Another practical issue is how to set the dimension d of the central subspace. We have proposed the use of BIC as a natural tool in our maximum likelihood setting. It provided good results but may be not suited when the sample size is too small. A more complete study and comparison with other solutions would be interesting.

To conclude, Student SIR shows good performance in the presence of outliers and is performing equally well in case of Gaussian errors. In our experiments, the algorithm has shown fast convergence being a promising alternative to SIR since nowadays most datasets include outliers. Future work would be to extend this setting to a multivariate response following the lead of [34, 35].

7. Appendix: Proofs

7.1. Proof of Proposition 1

The proof generalizes the proof of Proposition 6 in [9] to the generalized Student case. It comes from (7) that $\mathbf{X}_y$ follows a generalized Student distribution $\mathcal{S}_p(\boldsymbol{\mu}_y, \mathbf{V}, \alpha)$ where $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \mathbf{VBc}(y)$. Generalized Student distributions have similar properties to Gaussian distributions (see for instance section 5.5 in [16]). In particular any affine transformation of a generalized Student distribution remains in this family. It follows that $\mathbf{B}^T\mathbf{X}|Y = y$ is distributed as $\mathcal{S}_d(\mathbf{B}^T\boldsymbol{\mu}_y, \mathbf{B}^T\mathbf{VB}, \alpha)$. Similarly, marginals and conditional distributions are retained in the family. It follows that $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}, Y = y$ is also a generalized Student distribution $\mathcal{S}_p(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{V}}, \tilde{\alpha}, \tilde{\gamma})$ with

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \boldsymbol{\mu}_y + \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y) \\
&= \boldsymbol{\mu} + \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}) \\
\tilde{\mathbf{V}} &= \mathbf{V} - \mathbf{VB}(\mathbf{B}^T\mathbf{VB})^{-1}\mathbf{B}^T\mathbf{V} \\
\tilde{\alpha} &= \alpha + d \\
\tilde{\gamma} &= \frac{1}{2} + (\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y)^T(\mathbf{B}^T\mathbf{VB})^{-1}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\boldsymbol{\mu}_y) \\
&= \frac{1}{2} + \boldsymbol{\varepsilon}^T\mathbf{B}(\mathbf{B}^T\mathbf{VB})^{-1}\mathbf{B}^T\boldsymbol{\varepsilon} \ ,
\end{aligned}
$$

from which it is clear that $\tilde{\mathbf{V}}, \tilde{\alpha}, \tilde{\gamma}$ and $\tilde{\boldsymbol{\mu}}$ do not depend on y. It follows that $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}, Y = y$ has the same distribution as $\mathbf{X}|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ for all values $\mathbf{x}$. Consequently Y is independent on $\mathbf{X}$ conditionally to $\mathbf{B}^T\mathbf{X}$ which implies that $Y|\mathbf{X} = \mathbf{x}$ and $Y|\mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x}$ have identical distributions for all values $\mathbf{x}$. ∎

Note that for the proof of the proposition, it was necessary to show that the independence on y holds for each parameter of the distribution and not only for the mean. The independence on y of the mean is actually straightforward using [9] where it appears that the proof that $E[\mathbf{X}|\mathbf{B}^T\mathbf{X}, Y = y]$

does not depend on y is independent on the distribution of $\boldsymbol{\varepsilon}$. Indeed the proof uses only the properties of the conditional expectation seen as a projection operator. This means that in our case also, $\mathbf{B}$ corresponds to the *mean* central subspace as defined by $E[\mathbf{X}|\mathbf{B}^T\mathbf{X}, Y = y] = E[\mathbf{X}|\mathbf{B}^T\mathbf{X}]$.

7.2. Proof of Lemma 1

The proof is adapted from the proof of lemma 1 in [10] taking into account the additional quantities $\bar{u_i}$'s. Let us remark that

$$R \stackrel{def}{=} G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C}) - \log\det \mathbf{V} = \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_i^T\mathbf{V}^{-1}\mathbf{Z}_i, \tag{27}$$

where we have defined for $i = 1, \ldots, n$,

$$\begin{aligned}
\mathbf{Z}_i &= \boldsymbol{\mu} + \mathbf{V}\mathbf{B}\mathbf{C}^T\mathbf{s}_i - \mathbf{x}_i && (28)\\
&= (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}}) + \mathbf{V}\mathbf{B}\mathbf{C}^T(\mathbf{s}_i - \bar{\mathbf{s}}) - (\mathbf{x}_i - \bar{\mathbf{x}}) && (29)\\
&\stackrel{def}{=} \mathbf{Z}_1 + \mathbf{Z}_{2,i} - \mathbf{Z}_{3,i}. && (30)
\end{aligned}$$

Since $\mathbf{Z}_{2,.}$ and $\mathbf{Z}_{3,.}$ are centered, replacing the previous expansion in (27) yields

$R = \bar{u}\ \mathbf{Z}_1^T\mathbf{V}^{-1}\mathbf{Z}_1 + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{2,i} + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{3,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} - \frac{2}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i}$,

where

$$\begin{aligned}
\mathbf{Z}_1^T\mathbf{V}^{-1}\mathbf{Z}_1 &= (\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}})^T\mathbf{V}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}} + \mathbf{V}\mathbf{B}\mathbf{C}^T\bar{\mathbf{s}}),\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{2,i} &= tr(\mathbf{C}^T\mathbf{W}\mathbf{C}\mathbf{B}^T\mathbf{V}\mathbf{B}),\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{3,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} &= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathrm{tr}((\mathbf{x}_i - \bar{\mathbf{x}})^T\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}))\\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathrm{tr}(\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T)\\
&= \mathrm{tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma}) \text{ and}\\
\frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\mathbf{Z}_{2,i}^T\mathbf{V}^{-1}\mathbf{Z}_{3,i} &= tr(\mathbf{C}^T\mathbf{M}\mathbf{B}),
\end{aligned}$$

and the conclusion follows. ∎

7.3. Proof of Proposition 2

Cancelling the gradients of $G(\boldsymbol{\mu}, \mathbf{V}, \mathbf{B}, \mathbf{C})$ yields the system of equations

$$\frac{1}{2}\nabla_{\mu} G = \hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}} + \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}) = 0, \tag{31}$$

$$\frac{1}{2}\nabla_{B} G = \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T(\bar{u}\bar{\mathbf{s}}\bar{\mathbf{s}}^T + \mathbf{W})\hat{\mathbf{C}} - \mathbf{M}^T\hat{\mathbf{C}} + \bar{u}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})\bar{\mathbf{s}}^T\hat{\mathbf{C}} = 0, \tag{32}$$

$$\frac{1}{2}\nabla_{C} G = \bar{u}(\bar{\mathbf{s}}\bar{\mathbf{s}}^T\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} + \bar{\mathbf{s}}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})^T\hat{\mathbf{B}}) + \mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} - \mathbf{M}\hat{\mathbf{B}} = 0, \tag{33}$$

$$\nabla_{V} G = \hat{\mathbf{V}}^{-1} + \hat{\mathbf{B}}\hat{\mathbf{C}}^T(\bar{u}\bar{\mathbf{s}}\bar{\mathbf{s}}^T + \mathbf{W})\hat{\mathbf{C}}\hat{\mathbf{B}}^T + \tag{34}$$

$$- \hat{\mathbf{V}}^{-1}\left(\bar{u}(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})(\hat{\boldsymbol{\mu}} - \bar{\mathbf{x}})^T + \boldsymbol{\Sigma}\right)\hat{\mathbf{V}}^{-1} = 0. \tag{35}$$

From (31), we have

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\bar{\mathbf{s}}. \tag{36}$$

Replacing in (32) and (33) yields the simplified system of equations

$$\hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}) = \mathbf{M}^T\hat{\mathbf{C}}, \tag{37}$$

$$\mathbf{W}\hat{\mathbf{C}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}) = \mathbf{M}\hat{\mathbf{B}}. \tag{38}$$

It follows from the last equality that

$$\hat{\mathbf{C}} = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}\,. \tag{39}$$

Multiplying (37) by $\mathbf{B}^T\mathbf{V}\mathbf{B}$ on the left, we get

$$\hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} = \mathbf{M}^T\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}, \tag{40}$$

and assuming $\mathbf{W}$ is regular, (38) entails $\hat{\mathbf{C}}(\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}) = \mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}$. Replacing in (40) yields

$$\hat{\mathbf{V}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}} = \mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}. \tag{41}$$

Now, multiplying (34) on the left and on the right by $\hat{\mathbf{V}}$ and taking account of (36) entails

$$\boldsymbol{\Sigma} = \hat{\mathbf{V}} + \hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})\hat{\mathbf{B}}^T\hat{\mathbf{V}}. \tag{42}$$

As a consequence of (42), it comes

$$\boldsymbol{\Sigma}\hat{\mathbf{B}} = \hat{\mathbf{V}}\hat{\mathbf{B}}(\mathbf{I} + \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}), \tag{43}$$

and

$$\hat{\mathbf{V}}\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\mathbf{I} + \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}. \tag{44}$$

Using this expression of $\hat{\mathbf{V}}\hat{\mathbf{B}}$ above in (41), it comes

$$\hat{\mathbf{B}}\left(\mathbf{I} + (\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}})^{-1}\right)^{-1} = \mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}}\ , \tag{45}$$

which means that the columns of $\hat{\mathbf{B}}$ are stable by $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$ and thus are eigenvectors of $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$. Let us denote by $\lambda_1, \ldots, \lambda_d$ the associated eigenvalues. Matrix $\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}$ is of size $p \times p$ and of rank at most $min(h, p)$ since $\mathbf{W}$ is assumed to be regular. In practice we will assume $h \geq d$ and $p \geq d$. Therefore $d \leq min(h, p)$. It remains to show that $\lambda_1, \ldots, \lambda_d$ are the d largest eigenvalues. To this aim, we observe that using successively (38) and (42),

$$\begin{aligned} G(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) &= \log\det\hat{\mathbf{V}} + \mathrm{trace}(\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\hat{\mathbf{B}}\mathbf{C}^T\mathbf{W}) + \mathrm{trace}(\mathbf{V}^{-1}\mathbf{\Sigma}) - 2\mathrm{trace}(\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{M}) \\ &= \log\det\hat{\mathbf{V}} + \mathrm{trace}(\hat{\mathbf{M}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T) + p + \mathrm{trace}(\hat{\mathbf{M}}\hat{\mathbf{B}}\hat{\mathbf{C}}^T) - 2\mathrm{trace}(\hat{\mathbf{B}}\hat{\mathbf{C}}^T\mathbf{M}) \\ &= p + \log\det\hat{\mathbf{V}}. \end{aligned}$$

Let us consider the two following matrices, $\mathbf{\Delta}_1 = \mathbf{B}\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}$ and $\mathbf{\Delta}_2 = \hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\mathbf{B}$. $\mathbf{\Delta}_1$ is $p \times p$ of rank at most d and $\mathbf{\Delta}_2$ is $d \times d$ of rank d, invertible with positive eigenvalues denoted by $\delta_1, \ldots, \delta_d$. The eigenvalues of $\mathbf{\Delta}_2$ are that of $\mathbf{\Delta}_1$ too. Indeed consider $\mathbf{y}_k$ an eigenvector for δ_k, then $\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}}\hat{\mathbf{B}}^T\hat{\mathbf{V}}\mathbf{B}\mathbf{y}_k = \delta_k\mathbf{y}_k$. Multiplying on the left by $\hat{\mathbf{B}}$ and considering $\mathbf{z}_k = \hat{\mathbf{B}}\mathbf{y}_k$, it comes that δ_k is also an eigenvalue for $\mathbf{\Delta}_1$. Using (42), it follows then

$$\log\det\hat{\mathbf{V}} = \log\det\mathbf{\Sigma} - \log\det(\mathbf{I} + \mathbf{\Delta}_1) = \log\det\mathbf{\Sigma} - \sum_{k=1}^{d}\log(1 + \delta_k)\ .$$

Multiplying (45) by $\hat{\mathbf{B}}^T$ and using $\hat{\mathbf{B}}^T\hat{\mathbf{B}} = \mathbf{I}$, it comes
$\mathbf{I} + \mathbf{\Delta}_2^{-1} = (\hat{\mathbf{B}}^T\mathbf{\Sigma}^{-1}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M}\hat{\mathbf{B}})^{-1} = diag(1/\lambda_k)$ from which $\delta_k = \frac{1}{1-\lambda_k} - 1$ can be deduced. Finally,

$$G(\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) = p + \log\det\mathbf{\Sigma} + \sum_{k=1}^{d}\log(1 - \lambda_k)\ .$$

G is then minimized when the λ_k are the largest. As a consequence of (42), it also comes that

$$\hat{\mathbf{V}} = \mathbf{\Sigma} - \hat{\mathbf{V}}\hat{\mathbf{B}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})\hat{\mathbf{B}}^T\hat{\mathbf{V}}. \tag{46}$$

Replacing $\hat{\mathbf{V}}\hat{\mathbf{B}}$ in (46) by the expression given in (37), it comes

$$\hat{\mathbf{V}} = \mathbf{\Sigma} - \mathbf{M}^T\hat{\mathbf{C}}(\hat{\mathbf{C}}^T\mathbf{W}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^T\mathbf{M}. \tag{47}$$

Grouping the results in (47), (39), (36) and the considerations after (45) gives the Proposition. ■

7.4. Proof of Corollary 1.

Let us remark that, under (23), the coefficients W_{ij} of $\mathbf{W}$ have an explicit form:

$$\begin{aligned}
\mathbf{W} &= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{2}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\bar{\mathbf{s}}^T + \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \bar{\mathbf{s}}\bar{\mathbf{s}}^T \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{2ff^t}{\bar{u}} + \frac{ff^t}{\bar{u}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{u_i}\ \mathbf{s}_i\mathbf{s}_i^T - \frac{ff^t}{\bar{u}},
\end{aligned}$$

where $f = (f_1, ..., f_h)$. Using (23) the first sum corresponds to $\mathrm{diag}(f_1, ..., f_h)$ leading to $\mathbf{W} = \mathrm{diag}(f_1, ..., f_h) - \dfrac{ff^t}{\bar{u}}$. The inverse matrix of $\mathbf{W}$ can be calculated using Sherman-Morrison formula:

$$\mathbf{W}^{-1} = \mathrm{diag}\left(\frac{1}{f_1}, \ldots, \frac{1}{f_h}\right) + \frac{1}{f_{h+1}}\mathbf{O},$$

where $\mathbf{O}$ is the $h \times h$ matrix defined by $O_{ij} = 1$ for all $(i, j) \in \{1, \ldots, h\} \times \{1, \ldots, h\}$. Using (23) the jth row of $\mathbf{M}$ is given by:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\bar{u_i}(\mathbb{1}\{y_i \in S_j\} - \bar{s_j})(\mathbf{x}_i - \bar{\mathbf{x}})^T &= \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_j\}\mathbf{x}_i^T - \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_j\}\bar{\mathbf{x}}^T \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\bar{s_j}\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\bar{s_j}\bar{\mathbf{x}}^T \\ &= f_j\bar{\mathbf{x}_j}^T - f_j\bar{\mathbf{x}}^T - f_j\bar{\mathbf{x}}^T + f_j\bar{\mathbf{x}}^T \\ &= f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T,\end{aligned}$$

for all $j = 1, \ldots, h$. Now taking into account that $\mathbf{O}^2 = h\mathbf{O}$, we have

$$\begin{aligned}\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M} &= \sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T + \frac{1}{f_{h+1}}\mathbf{M}^T\mathbf{O}\mathbf{M} \\ &= \sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T + \frac{1}{hf_{h+1}}(\mathbf{M}^T\mathbf{O})(\mathbf{M}^T\mathbf{O})^T. \qquad (48)\end{aligned}$$

Now, remarking that all the columns of $\mathbf{M}^T\mathbf{O}$ are equal to

$$\sum_{j=1}^{h} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) = \sum_{j=1}^{h+1} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) - f_{h+1}(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}}) = -f_{h+1}(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}}),$$

where $f_{h+1} = \frac{1}{n}\sum_{i=1}^{n}\bar{u_i}\mathbb{1}\{y_i \in S_{h+1}\} = \bar{u} - \sum_{j=1}^{h} f_j$ it follows that

$$(\mathbf{M}^T\mathbf{O})(\mathbf{M}^T\mathbf{O})^T = hf_{h+1}^2(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{h+1} - \bar{\mathbf{x}})^T$$

and thus replacing in (48) yields

$$\mathbf{M}^T\mathbf{W}^{-1}\mathbf{M} = \sum_{j=1}^{h+1} f_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T = \mathbf{\Gamma}.$$

The result is then a consequence of Proposition 2. ■

Acknowledgments

The authors would like to thank Didier Fraix-Burnet for his contribution to the data. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

[1] R. D. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (435) (1996) 983–992.

[2] X. Yin, B. Li, R. D. Cook, Successive direction extraction for estimating the central subspace in a multiple-index regression, Journal of Multivariate Analysis 99 (8) (2008) 1733–1757.

[3] K.-C. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (414) (1991) 316–327.

[4] P. Hall, K.-C. Li, On almost linearity of low dimensional projections from high dimensional data, The Annals of Statistics 21 (2) (1993) 867–889.

[5] K. Fukumizu, F. R. Bach, M. I. Jordan, Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, The Journal of Machine Learning Research 5 (2004) 73–99.

[6] B. Li, Y. Dong, Dimension reduction for nonelliptically distributed predictors, The Annals of Statistics 37 (3) (2009) 1272–1298.

[7] K. Fukumizu, F. R. Bach, M. I. Jordan, Kernel dimension reduction in regression, The Annals of Statistics 37 (4) (2009) 1871–1905.

[8] R. D. Cook, S. Weisberg, Sliced inverse regression for dimension reduction: Comment, Journal of the American Statistical Association 86 (414) (1991) 328–332.

[9] R. D. Cook, Fisher lecture: Dimension reduction in regression, Statistical Science 22 (1) (2007) 1–26.

[10] C. Bernard-Michel, L. Gardes, S. Girard, Gaussian regularized sliced inverse regression, Statistics and Computing 19 (1) (2009) 85–98.

[11] E. Bura, R. D. Cook, Extending sliced inverse regression: The weighted chi-squared test, Journal of the American Statistical Association 96 (455) (2001) 996–1003.

[12] S. J. Sheather, J. W. McKean, A comparison of procedures based on inverse regression, Lecture Notes-Monograph Series 31 (1997) 271–278.

[13] U. Gather, T. Hilker, C. Becker, A note on outlier sensitivity of sliced inverse regression, Statistics: A Journal of Theoretical and Applied Statistics 36 (4) (2002) 271–281.

[14] U. Gather, T. Hilker, C. Becker, A robustified version of sliced inverse regression, in: Statistics in Genetics and in the Environmental Sciences, Trends in Mathematics, Springer, 2001, Ch. 2, pp. 147–157.

[15] Y. Dong, Z. Yu, L. Zhu, Robust inverse regression for dimension reduction, Journal of Multivariate Analysis 134 (2015) 71–81.

[16] S. Kotz, S. Nadarajah, Multivariate t Distributions and their Applications, Cambridge, 2004.

[17] G. McLachlan, D. Peel, Robust mixture modelling using the t distribution, Statistics and Computing 10 (2000) 339–348.

[18] D. F. Andrews, C. L. Mallows, Scale mixtures of normal distributions, Journal of the Royal Statistical Society. Series B (Methodological) 36 (1) (1974) 99–102.

[19] N. L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, vol.2, 2nd edition, John Wiley & Sons, New York, 1994.

[20] L. Ferré, Determining the dimension in sliced inverse regression and related methods, Journal of the American Statistical Association 93 (441) (1998) 132–140.

[21] M. Chavent, S. Girard, V. Kuentz-Simonet, B. Liquet, T. M. N. Nguyen, J. Saracco, A sliced inverse regression approach for data stream, Computational Statistics 29 (5) (2014) 1129–1152.

[22] B. Liquet, J. Saracco, A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches, Computational Statistics 27 (1) (2012) 103–125.

[23] J. R. Schott, Determining the Dimensionality in Sliced Inverse Regression, Journal of the American Statistical Association 89 (425) (1994) 141–148.

[24] S. Velilla, Assessing the number of linear components in a general regression problem, Journal of the American Statistical Association 93 (443) (1998) 1088–1098.

[25] M. P. Barrios, S. Velilla, A bootstrap method for assessing the dimension of a general regression problem, Statistics & Probability Letters 77 (3) (2007) 247–255.

[26] L. Zhu, B. Miao, H. Peng, On Sliced Inverse Regression With High-Dimensional Covariates, Journal of the American Statistical Association 101 (2006) 630–643.

[27] R. E. Kass, A. E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995) 773–795.

[28] C. Giraud, Introduction to high-dimensional statistics, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, 2014.

[29] H. Wang, L. Ni, C.-L. Tsai, Improving dimension reduction via contour-projection, Statistica Sinica 18 (1) (2008) 299–311.

[30] R. Luo, H. Wang, C.-L. Tsai, Contour projected dimension reduction, The Annals of Statistics 37 (6B) (2009) 3743–3778.

[31] J. Zhou, Robust dimension reduction based on canonical correlation, Journal of Multivariate Analysis 100 (1) (2009) 195–209.

[32] R. D. Cook, L. Forzani, D. R. Tomassi, Ldr: a package for likelihood-based sufficient dimension reduction, Journal of Statistical Software 39 (3) (2011) 1–20.

[33] A. Chiancone, S. Girard, J. Chanussot, Collaborative sliced inverse regression, Communications in Statistics - Theory and Methods. To appear.

[34] L. Barreda, A. Gannoun, J. Saracco, Some extensions of multivariate SIR, Journal of Statistical Computation and Simulation 77 (2007) 1–17.

[35] R. Coudret, S. Girard, J. Saracco, A new sliced inverse regression method for multivariate response, Computational Statistics and Data Analysis 77 (2014) 285–299.